Ejercicio 1: Especificación y Análisis de Regresión

Michigan Schools - Desempeño en Matemáticas

Pregunta i: Variable Proxy para Pobreza

¿Por qué lnchprg es una proxy razonable para poverty?

Respuesta:

La variable `lnchprg` (porcentaje de estudiantes elegibles para el programa federal de comidas escolares subvencionadas) es una **proxy razonable** para `poverty` por las siguientes razones:

1. **Criterio de elegibilidad basado en ingreso:** Los estudiantes son elegibles para el programa de comidas subsidiadas si su familia cumple con ciertos umbrales de pobreza establecidos por el gobierno federal. Específicamente, familias con ingresos por debajo del 130 % de la línea de pobreza califican para comidas gratuitas, y aquellas entre 130 % y 185 % califican para comidas a precio reducido.

2. **Alta correlación con pobreza:** Existe una correlación directa y fuerte entre el porcentaje de estudiantes elegibles y el nivel de pobreza en la comunidad escolar:

$$\text{Corr}(\text{lnchprg}, \text{poverty}) \approx 1 \tag{1}$$

3. **Medición objetiva y disponible:** A diferencia de medidas directas de pobreza que pueden ser difíciles de obtener a nivel escolar, `lnchprg` se registra administrativamente y está ampliamente disponible en bases de datos educativas.

4. **Captura efectos socioeconómicos relevantes:** La variable captura múltiples dimensiones de desventaja socioeconómica que afectan el desempeño académico:

   - Recursos limitados en el hogar
   - Estrés financiero familiar
   - Acceso limitado a recursos educativos complementarios
   - Menor capital cultural

Formalmente, si denotamos la variable no observada como poverty^*, entonces `lnchprg` es una buena proxy si:

$$\text{lnchprg} = \text{poverty}^* + \eta \tag{2}$$

donde η es un error de medición con $E[\eta|X] = 0$ y $\text{Cov}(\text{lnchprg}, \text{poverty}^*) \gg 0$.

Pregunta ii.a: Cambio en el Coeficiente de log(expend)

¿Por qué el efecto del gasto es menor en columna (2)?

Explicación del cambio:

El coeficiente de log(expend) disminuye de **11.13** en el modelo (1) a **7.75** en el modelo (2). Este cambio se debe al **sesgo por variable omitida (OVB)** que estaba presente en el modelo (1).

Análisis formal del sesgo:

En el modelo (1) sin `lnchprg`:

$$\text{math10} = \beta_0 + \beta_1 \log(\text{expend}) + \beta_2 \log(\text{enroll}) + u_1 \tag{3}$$

donde $u_1 = \beta_3 \text{poverty} + \varepsilon$ contiene la variable omitida.

El sesgo en $\hat{\beta}_1$ está dado por:

$$\text{Sesgo}(\hat{\beta}_1) = \beta_3 \cdot \frac{\text{Cov}(\log(\text{expend}), \text{poverty})}{\text{Var}(\log(\text{expend}))} \tag{4}$$

Análisis de signos:

- $\beta_3 < 0$: Mayor pobreza $\rightarrow$ menor desempeño (efecto negativo)
- $\text{Cov}(\log(\text{expend}), \text{poverty}) < 0$: Escuelas con mayor gasto tienden a servir poblaciones menos pobres (distritos ricos gastan más)

Por lo tanto:

$$\text{Sesgo}(\hat{\beta}_1) = (\text{negativo}) \times \frac{(\text{negativo})}{(\text{positivo})} = \text{POSITIVO} \tag{5}$$

Conclusión: El modelo (1) sobreestima el efecto del gasto porque parte del efecto observado se debe realmente a diferencias en composición socioeconómica.

¿Sigue siendo estadísticamente significativo?

Prueba de hipótesis:

$$H_0 : \beta_1 = 0 \quad \text{(el gasto no tiene efecto)} \tag{6}$$

$$H_1 : \beta_1 \neq 0 \quad \text{(el gasto tiene efecto)} \tag{7}$$

Estadístico t:

$$t = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{7{,}75}{3{,}04} \approx 2{,}55 \tag{8}$$

Valor crítico: Con $n - k - 1 = 428 - 3 - 1 = 424$ grados de libertad:

$$t_{0{,}025,424} \approx 1{,}96 \tag{9}$$

Decisión: Como $|t| = 2{,}55 > 1{,}96$, **rechazamos** H_0 al nivel de significancia del 5 %.

Conclusión: SÍ, el efecto sigue siendo estadísticamente significativo y positivo. Un aumento del 10 % en el gasto por estudiante se asocia con un aumento de aproximadamente 0.775 puntos porcentuales en la tasa de aprobación, manteniendo constantes la matrícula y el nivel de pobreza.

Pregunta ii.b: Cambio de Signo en log(enroll)

¿Por qué cambia el signo de log(enroll)?

El coeficiente de log(enroll) cambia de **+0.022** (positivo, no significativo) en el modelo (1) a **-1.26** (negativo, significativo) en el modelo (2).

Explicación mediante sesgo por variable omitida:

El sesgo en $\hat{\beta}_2$ (coeficiente de enrollment) es:

$$\text{Sesgo}(\hat{\beta}_2) = \beta_3 \cdot \frac{\text{Cov}(\log(\text{enroll}), \text{poverty})}{\text{Var}(\log(\text{enroll}))} \tag{10}$$

Relaciones empíricas:

- $\beta_3 < 0$: Mayor pobreza reduce el desempeño
- $\text{Cov}(\log(\text{enroll}), \text{poverty}) > 0$: Escuelas más grandes tienden a estar en áreas urbanas con mayor concentración de pobreza

Por lo tanto:

$$\text{Sesgo}(\hat{\beta}_2) = (\text{negativo}) \times \frac{(\text{positivo})}{(\text{positivo})} = \text{NEGATIVO} \tag{11}$$

Pero en este caso, el sesgo es *hacia arriba* (hacia cero o positivo), lo que significa que el modelo (1) oculta el efecto negativo real del tamaño escolar.

Interpretación del fenómeno:

En el modelo (1), el coeficiente positivo (aunque no significativo) capturaba una **correlación espuria**:

- Escuelas grandes → áreas urbanas → mayor pobreza → menor desempeño
- Esta cadena causal se confundía con el efecto directo del tamaño

Al controlar por `lnchprg` en el modelo (2), **aislamos el efecto genuino** del tamaño escolar, revelando que:

$$\frac{\partial \text{math10}}{\partial \log(\text{enroll})} < 0 \quad \text{(manteniendo pobreza constante)} \tag{12}$$

Pregunta ii.c: Interpretación del Efecto de Matrícula

Tasas de aprobación menores en escuelas más grandes

El coeficiente $\hat{\beta}_2 = -1{,}26$ en el modelo (2) tiene la siguiente interpretación:
Interpretación semi-elasticidad:

$$\Delta \text{math10} = \beta_2 \times \Delta \log(\text{enroll}) \times 100 \tag{13}$$

Ejemplo concreto: Un aumento del 10 % en la matrícula ($\Delta \log(\text{enroll}) = 0{,}10$) se asocia con:

$$\Delta \text{math10} = -1{,}26 \times 0{,}10 = -0{,}126 \text{ puntos porcentuales} \tag{14}$$

O equivalentemente, duplicar la matrícula ($\Delta \log(\text{enroll}) = \log(2) \approx 0{,}693$):

$$\Delta \text{math10} = -1{,}26 \times 0{,}693 \approx -0{,}87 \text{ puntos porcentuales} \tag{15}$$

¿Por qué ocurre esto?
Posibles mecanismos causales (manteniendo constantes gasto y pobreza):

1. **Dilución de recursos por estudiante:** Aunque el gasto total sea similar, escuelas grandes pueden tener:
   - Salones más grandes (menos atención individual)
   - Administración más burocrática
   - Menor comunidad/cohesión escolar
2. **Efectos de pares negativos:** En escuelas grandes, los estudiantes pueden estar más expuestos a:
   - Mayor heterogeneidad (más difícil enseñar)
   - Ambiente menos personalizado
   - Menor monitoreo individual
3. **Economías de escala decrecientes:** Más allá de cierto tamaño, los beneficios de escala se agotan y aparecen costos de coordinación.

Nota importante: Este es un efecto *ceteris paribus* (manteniendo constantes gasto y pobreza). No significa que todas las escuelas grandes sean malas, sino que, comparando escuelas con similar gasto y composición socioeconómica, las más pequeñas tienden a tener mejor desempeño.

Pregunta ii.d: Interpretación del Coeficiente de lnchprg

Efecto de la variable proxy de pobreza

El coeficiente $\hat{\beta}_3 = -0{,}324$ representa el efecto del porcentaje de estudiantes en el programa de comidas subsidiadas.

Interpretación directa:

$$\frac{\partial \text{math10}}{\partial \text{lnchprg}} = -0{,}324 \tag{16}$$

Esto significa: **Un aumento de un punto porcentual** en el porcentaje de estudiantes elegibles para comidas subsidiadas se asocia con una **disminución de 0.324 puntos porcentuales** en la tasa de aprobación del examen de matemáticas, manteniendo constantes el gasto y la matrícula.

Ejemplos numéricos:

- Si `lnchprg` aumenta de 20 % a 30 % (incremento de 10 puntos):

$$\Delta \text{math10} = -0{,}324 \times 10 = -3{,}24 \text{ puntos porcentuales} \tag{17}$$

- Comparando dos escuelas idénticas en gasto y tamaño, pero donde una tiene 50 % de estudiantes en el programa y la otra 25 %:

$$\Delta \text{math10} = -0{,}324 \times (50 - 25) = -8{,}1 \text{ puntos porcentuales} \tag{18}$$

Significancia del efecto:
Error estándar: $\text{SE}(\hat{\beta}_3) = 0{,}036$
Estadístico t:

$$t = \frac{-0{,}324}{0{,}036} = -9{,}0 \tag{19}$$

Con $|t| = 9{,}0 \gg 1{,}96$, este efecto es **altamente significativo** estadísticamente y también **sustancialmente importante** en magnitud.

Interpretación económica:
Este coeficiente captura el efecto de múltiples canales de desventaja socioeconómica:

- **Recursos en el hogar:** Familias de bajos ingresos tienen menos recursos para apoyar el aprendizaje (libros, computadoras, espacio de estudio, tutoría)
- **Capital humano parental:** Correlación entre pobreza y educación parental
- **Estrés y salud:** Inseguridad alimentaria y estrés financiero afectan concentración y desarrollo cognitivo
- **Efectos de pares:** Alta concentración de pobreza puede generar efectos de pares negativos

Pregunta ii.e: Incremento Sustancial en R^2

¿Qué opina del incremento en R^2 de 0.0297 a 0.1893?

Magnitud del cambio:

$$R^2_{\text{Modelo 1}} = 0{,}0297 \tag{20}$$

$$R^2_{\text{Modelo 2}} = 0{,}1893 \tag{21}$$

$$\Delta R^2 = 0{,}1596 \text{ (incremento absoluto)} \tag{22}$$

$$\text{Incremento relativo} = \frac{0{,}1596}{0{,}0297} \times 100\,\% = 537\,\% \tag{23}$$

Interpretación:

1. **Poder explicativo:**
   - Modelo 1: Solo explica 2.97 % de la variación en `math10`
   - Modelo 2: Explica 18.93 % de la variación
   - La variable `lnchprg` por sí sola añade ≈ 16 puntos porcentuales de poder explicativo
2. **Importancia de la composición socioeconómica:**

   Este incremento dramático revela que **la pobreza es el factor más importante** para explicar el desempeño en matemáticas, más que el gasto o el tamaño de la escuela.

   Formalmente, podemos descomponer la varianza explicada:

   $$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)} \tag{24}$$

   La contribución marginal de `lnchprg`:

   $$R^2_{\text{lnchprg — otros}} = R^2_{\text{completo}} - R^2_{\text{sin lnchprg}} = 0{,}1596 \tag{25}$$

3. **Validez del modelo:**

   El incremento sustancial confirma que:
   - Había un **sesgo severo por variable omitida** en el modelo (1)
   - `lnchprg` es una proxy **relevante y poderosa**
   - El modelo (2) está **mejor especificado**
4. **Aún queda varianza sin explicar:**

   Incluso con $R^2 = 0{,}1893$, todavía hay $1 - 0{,}1893 = 81\,\%$ de variación sin explicar. Esto sugiere que otros factores también importan:

- Calidad docente
- Prácticas pedagógicas
- Recursos del hogar no capturados por pobreza
- Motivación estudiantil
- Capital social de la comunidad

Advertencia sobre R^2:
Un R^2 alto no implica necesariamente un buen modelo para:

- **Inferencia causal:** Un R^2 alto no garantiza que los coeficientes sean insesgados
- **Predicción fuera de muestra:** Podría haber sobreajuste
- **Relevancia política:** Los efectos individuales importan más que el R^2 total

Sin embargo, en este contexto, el incremento es **evidencia sólida** de que omitir `lnchprg` generaba un problema serio de especificación.

Pregunta iii: Prueba de Especificación

Comparación entre formas funcionales

Queremos comparar:
Especificación A (logarítmica):

$$\text{math10} = \beta_0 + \beta_1 \log(\text{expend}) + \beta_2 \log(\text{enroll}) + \beta_3 \text{lnchprg} + u \tag{26}$$

Especificación B (niveles y cuadrados):

$$\text{math10} = \gamma_0 + \gamma_1 \text{expend} + \gamma_2 \text{expend}^2 + \gamma_3 \text{enroll} + \gamma_4 \text{enroll}^2 + \gamma_5 \text{lnchprg} + v \tag{27}$$

Procedimiento formal de prueba

Método 1: Prueba de Davidson-MacKinnon (J-Test)

Paso 1: Estimar ambos modelos y obtener valores predichos:

$$\hat{y}_A = \text{predicciones del modelo A (logarítmico)} \tag{28}$$
$$\hat{y}_B = \text{predicciones del modelo B (niveles)} \tag{29}$$

Paso 2: Probar si el modelo A anida al modelo B:
Estimar el modelo aumentado:

$$\text{math10} = \beta_0 + \beta_1 \log(\text{expend}) + \beta_2 \log(\text{enroll}) + \beta_3 \text{lnchprg} + \delta_1 \hat{y}_B + \text{error} \tag{30}$$

Hipótesis:

$$H_0 : \delta_1 = 0 \quad \text{(modelo A es correcto)} \tag{31}$$
$$H_1 : \delta_1 \neq 0 \quad \text{(modelo A es inadecuado)} \tag{32}$$

Estadístico de prueba:

$$t = \frac{\hat{\delta}_1}{\text{SE}(\hat{\delta}_1)} \sim t_{n-k-1} \tag{33}$$

Paso 3: Probar si el modelo B anida al modelo A:
Estimar:

$$\begin{aligned}\text{math10} = \gamma_0 + \gamma_1 \text{expend} + \gamma_2 \text{expend}^2 + \gamma_3 \text{enroll} \\ + \gamma_4 \text{enroll}^2 + \gamma_5 \text{lnchprg} + \delta_2 \hat{y}_A + \text{error}\end{aligned} \tag{34}$$

Hipótesis:

$$H_0 : \delta_2 = 0 \quad \text{(modelo B es correcto)} \tag{35}$$
$$H_1 : \delta_2 \neq 0 \quad \text{(modelo B es inadecuado)} \tag{36}$$

Decisión:

H_0: Modelo A	H_0: Modelo B	**Conclusión**
No rechazar	Rechazar	Preferir modelo A
Rechazar	No rechazar	Preferir modelo B
No rechazar	No rechazar	Ambos adecuados (elegir por parsimonia)
Rechazar	Rechazar	Ninguno adecuado (buscar nueva especificación)

Método 2: Criterios de Información

Criterio de Información de Akaike (AIC):

$$\text{AIC} = -2\log(L) + 2k \tag{37}$$

donde L es la verosimilitud y k es el número de parámetros.
Criterio Bayesiano de Información (BIC):

$$\text{BIC} = -2\log(L) + k\log(n) \tag{38}$$

donde n es el tamaño de muestra.
Regla de decisión: Elegir el modelo con **menor** AIC o BIC.
Ventaja del BIC: Penaliza más fuertemente la complejidad del modelo, favoreciendo la parsimonia.

Método 3: Test RESET de Ramsey

Idea: Probar si términos no lineales de $\hat{y}$ son significativos.
Para el modelo A:

$$\text{math10} = \beta_0 + \beta_1 \log(\text{expend}) + \beta_2 \log(\text{enroll}) + \beta_3 \text{lnchprg} + \alpha_1 \hat{y}^2 + \alpha_2 \hat{y}^3 + \text{error} \quad (39)$$

Hipótesis:

$$H_0 : \alpha_1 = \alpha_2 = 0 \quad \text{(especificación correcta)} \quad (40)$$
$$H_1 : \text{al menos uno} \neq 0 \quad \text{(mala especificación)} \quad (41)$$

Estadístico F:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k)} \sim F_{q,n-k} \quad (42)$$

donde SSR_r es la suma de residuos al cuadrado del modelo restringido (sin $\hat{y}^2, \hat{y}^3$) y SSR_{ur} del no restringido.

Método 4: Validación Cruzada

Procedimiento:

1. Dividir datos en conjunto de entrenamiento (70 %) y prueba (30 %)
2. Estimar ambos modelos en datos de entrenamiento
3. Predecir en datos de prueba
4. Calcular error cuadrático medio (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} (y_i - \hat{y}_i)^2} \quad (43)$$

5. Elegir modelo con menor RMSE

Ventaja: Mide capacidad predictiva real fuera de muestra.

Método 5: Test F para Restricciones

Si queremos probar formalmente si la forma logarítmica es preferible, podemos usar el modelo más general que anida ambos:

$$\begin{aligned}\text{math10} = \beta_0 + \beta_1 \log(\text{expend}) + \beta_2 \log(\text{enroll}) + \beta_3 \text{lnchprg} \\ + \gamma_1 \text{expend} + \gamma_2 \text{expend}^2 + \gamma_3 \text{enroll} + \gamma_4 \text{enroll}^2 + u\end{aligned} \quad (44)$$

Probar especificación logarítmica:

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0 \tag{45}$$

$$H_1 : \text{al menos uno} \neq 0 \tag{46}$$

Estadístico F:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-q)} \sim F_{q,n-k-q} \tag{47}$$

con $q = 4$ restricciones.

Decisión: Si no rechazamos H_0, la forma logarítmica es adecuada (principio de parsimonia).

Resumen de Pasos Formales

1. **Estimar ambos modelos** y guardar:
   - Coeficientes estimados
   - Valores predichos
   - R^2, SSR
   - AIC, BIC
2. **Aplicar J-Test:**
   - Agregar $\hat{y}_B$ al modelo A y probar significancia
   - Agregar $\hat{y}_A$ al modelo B y probar significancia
   - Interpretar resultados según tabla de decisión
3. **Comparar criterios de información:**
   - Calcular AIC y BIC para ambos
   - Modelo con menor AIC/BIC es preferible
4. **Aplicar RESET:**
   - Probar cada modelo por separado
   - Si RESET rechaza, hay evidencia de mala especificación
5. **Validación cruzada:**
   - Dividir muestra
   - Estimar y predecir
   - Comparar RMSE
6. **Decisión final:**
   - Integrar evidencia de todos los tests

- Considerar teoría económica
- Considerar interpretabilidad
- Elegir modelo más parsimonioso si ambos son adecuados

Nota importante: En econometría, ninguna prueba es definitiva. La elección de especificación debe basarse en:

- Teoría económica
- Evidencia estadística múltiple
- Interpretabilidad
- Robustez de los resultados